A Logic for Checking the Probabilistic Steady-State Properties of Reaction Networks

Vincent Picard and **Anne Siegel** and **Jérémie Bourdon**
Université de Rennes 1
CNRS
Université de Nantes
Rennes/Nantes, France
vincent.picard@ens-cachan.org

Abstract

Designing probabilistic reaction models and determining their stochastic kinetic parameters are major issues in systems biology. In order to assist in the construction of reaction network models, we introduce a logic that allows one to express asymptotic properties about the steady-state stochastic dynamics of a reaction network. Basically, the formulas can express properties on expectancies, variances and co-variances. If a formula encoding for experimental observations on the system is not satisfiable then the reaction network model can be rejected. We demonstrate that deciding the satisfiability of a formula is NP-hard but we provide a decision method based on solving systems of polynomial constraints. We illustrate our method on a toy example.

1 Introduction

The dynamical quantitative analysis of systems of coupled chemical reactions also known as *reaction networks* is a major topic of interest in systems biology. Two main mathematical frameworks have been introduced to investigate their kinetic behavior [Helms, 2008]: ordinary differential equations at the population level and stochastic modeling at the single-cell level.

Ordinary differential equations (ODEs) provide deterministic trajectories for the average quantities of molecules at the population level. The time evolution of the quantities of molecules $\vec{x}$ is described by a system of ordinary differential equations of type $\frac{d\vec{x}}{dt} = S\vec{f}(\vec{x})$ where S is the stoichiometry matrix of the system and $\vec{f}$ is a vector of *fluxes* that depends on the current matter quantities. Usually the value of $\vec{f}$ is given by the law of mass actions although other laws may be used (Michaelis-Menten, Droop, ...). When all molecular species, reactions, kinetic laws and their parameters are known, numerical analysis algorithms allow one to compute approximate trajectories of the average quantities of molecules. When the system is either too large or not enough provided with experimental data, an alternative method is to consider the *steady-state* of the system, where the reactant concentrations are assumed to be constant because their production and consumption are balanced. In this case, the fluxes which depend on matter quantities are constant and must satisfy the equation $S\vec{f} = \vec{0}$. Based on the information provided by the stoichiometry matrix, *constraint-based approaches* allow finding the appropriate f subjected to $S\vec{f} = \vec{0}$ together with additional biological constraints in the *flux balance analysis* framework (FBA) [Orth *et al.*, 2010].

Among numerous applications, fluxes based methods can be used for *model validation and comparison*, that is, deciding whether a proposed set of reactions is consistent with the observed data or not. Here the observations consist of measuring some steady-state production rates of output metabolites, that is chemical species that are not consumed by the reactions. Hence we consider a steady-state where all metabolites quantities are constant, except the output metabolites. As an example consider the following reaction network

$$A \to B + C$$
$$B \to A + D$$

The production rates τ_C and τ_D of the output metabolites C and D can be derived from the fluxes which satisfy:

$$\frac{d\vec{x}}{dt} = S^\top \vec{f}(\vec{x}) = S^\top \vec{f} = \begin{pmatrix} 0 \\ 0 \\ \tau_C \\ \tau_D \end{pmatrix}. \quad (1)$$

Experimental observations on τ_C and τ_D can be encoded into a logical formula of type $\varphi = (\tau_C > 2\tau_D)$. Using equation (1), one can determine if φ is compatible with a given flux $\vec{f}$. This can be logically formalized as: a flux $\vec{f}$ is a model of φ ($\vec{f} \models \varphi$) if it satisfies both equations (1) and φ. If there is no value for $\vec{f}$ such that $\vec{f} \models \varphi$, in other words if φ is not satisfiable, then the reaction network can be rejected based on the data at hand. In the proposed example, it can be quickly checked that $\varphi = \tau_C > 2\tau_D$ is not satisfiable, so the data would refute the proposed reaction network. This type of reasoning is very useful for biologists who can eliminate modeling hypotheses based only on output slopes measurements and by checking the satisfiability of a formula. Notice that no information about kinetic laws or parameters have been used.

However, there exists some situations where constraints on steady-states fluxes are not sufficient to discriminate models.

As a toy illustrative example, let us consider the two following reaction models

$$(model\,1)\ \begin{cases} R_1: & A & \rightarrow & B+C \\ R_2: & B & \rightarrow & A+2D \end{cases} \quad (2)$$

$$(model\,2)\ \begin{cases} R_1: & A & \rightarrow & B+C+2D \\ R_2: & B & \rightarrow & A \end{cases} \quad (3)$$

A flux-based approach indicates that in both models, every flux $\vec{f} = (f_1, f_2)$ with $f_1 = f_2$ satisfies the balance constraints and that the accumulation rate of C equals f_1 whereas the accumulation rate of D equals $2f_1$. Thus both systems are equivalent from the point of view of fluxes approaches based on the average production rates of C and D. However, the steady-states of these models actually *can* be distinguished from each other. Intuitively, in model 1, the quantities of C and D should be negatively correlated while they should be positively correlated in model 2. To formalize this intuition, we need to focus on the single cell level, where stochastic fluctuations exist. This can be seen in Fig. 1 where individual trajectories of the system at the stochastic level for both models are depicted. It appears that the mean and variance quantities do not allow distinguishing models 1 and 2, whereas the covariance line (dark line) is clearly distinct between model 1 and 2.

Therefore, probabilistic modeling associated with stochastic data can be relevant to assist in the design of reaction networks. Moreover, the importance of dynamical stochastic modeling is continuously growing [Wilkinson, 2009] as biology intrinsically exhibits stochastic behaviors [McAdams and Arkin, 1997; Arkin *et al.*, 1998] and techniques of single-cell observations are improving. The reference method in stochastic modeling is to use the Gillespie stochastic simulation algorithm [Gillespie, 1976; 2007] that generates stochastic trajectories of a reaction network. The distribution of the sampled trajectories is solution to the Chemical Master Equation, which is the probabilistic equivalent of the law of mass actions. However using the Gillespie algorithm is computationally intensive and requires the knowledge of all reaction kinetic parameters as for the differential methods. Consequently, we develop in this work a logic which focuses on the stochastic steady-state properties and does not require information about the kinetic parameters.

Objective The aim of this article is to define a logic that permits to express the steady-state stochastic properties (means, variances, covariances) of the outputs, instead of only average production rates. In doing so, we would increase the rejection power of the fluxes based method by taking into account the fluctuations of individual cells. In section 2 we define the syntax and the semantics of the logic, in particular we define the satisfiability of formulas. In section 3 we demonstrate that deciding the satisfiability is NP-hard and we propose an algorithm to decide the satisfiability. In the last section 4 we apply these results on the introductory example.

2 Syntax and Semantics

Reaction networks We consider systems of chemical reactions known as reaction networks. A *reaction network* consists of n molecular species $X_1, \ldots, X_n$ that are involved in m chemical reactions $R_i : a_{i,1}X_1 + \cdots + a_{i,n}X_n \rightarrow b_{i,1}X_1 + \cdots + b_{i,n}X_n \quad (1 \leq i \leq m)$. The parameters $a_{i,j}$, $b_{i,j} \in \mathbb{N}$ are the *stoichiometry coefficients* of the reaction network. The number $a_{i,j}$ represents the quantity of X_j molecules consumed by the reaction R_i and the number $b_{i,j}$ represents the quantity of X_j molecules produced by the reaction R_j. The global effect of the reactions on the molecular quantities is often summarized by the *stoichiometry matrix* $S = (s_{i,j})_{1\leq i\leq m,\, 1\leq j\leq n}$ where $s_{i,j} = b_{i,j} - a_{i,j}$. In our notations, each row of the stoichiometry matrix represents the effect of one reaction on the quantity of molecules.
In this article we consider *discrete-time* dynamics of reaction networks. We denote $(\vec{x}_k)_{k\in\mathbb{N}}$ the discrete time stochastic process describing the number of molecules of each chemical specie at time k. For instance $(\vec{x}_k)_{k\in\mathbb{N}}$ may be generated by a discrete-time version of the Gillespie algorithm [Sandmann, 2008].

2.1 Syntax

Definition of terms We want to formally define the syntax of formulas describing some asymptotic properties on $(\vec{x}_k)_{k\in\mathbb{N}}$. More precisely we want to compare asymptotically polynomial expressions involving the first and second moments of $(\vec{x}_k)$. These polynomial expressions are the *terms* of our logic. We denote by $\mathcal{C} = \{X_1, \ldots, X_n\}$ the non-empty finite set of chemical species symbols. The algebra of *terms* is defined by structural induction as the *least* set $\mathcal{T}$ satisfying:

$$\forall X, Y \in \mathcal{C},\ \mathrm{Exp}(X) \in \mathcal{T},\ \mathrm{Var}(X) \in \mathcal{T},\ \mathrm{Cov}(X,Y) \in \mathcal{T},$$

$$\forall \lambda \in \mathbb{Q},\ \forall T_1, T_2 \in \mathcal{T},\ \lambda \in \mathcal{T},\ (\lambda \cdot T_1) \in \mathcal{T}, (T_1 + T_2) \in \mathcal{T}, (T_1 \times T_2) \in \mathcal{T}.$$

For the moment, Exp, Var and Cov are just function symbols, their semantics is defined later.

Example 1. $(\mathrm{Var}(X_1) + \mathrm{Cov}(X_3, X_4))$ *and* $((3 \cdot \mathrm{Exp}(X_1)) \times \mathrm{Var}(X_2))$ *are terms.*

Definition of formulas We are now able to define the syntax of the formulas which are used to compare two terms, that is two polynomial expressions involving the first and second moments of $(\vec{x}_k)_k$. In order to provide a simple definition, the only *atomic formulas* we introduce are the comparisons with 0:

$$\text{atomic formulas } : \mathcal{AF} = \{(T \geq 0) \,/\, T \in \mathcal{T}\}.$$

The formulas are atomic propositions connected with the classical logical operators. Formally, the set of *formulas* is defined by structural induction as the *least* set $\mathcal{F}$ satisfying: $\mathcal{AF} \subset \mathcal{F}$ and $\forall F_1, F_2 \in \mathcal{F}, \neg F_1 \in \mathcal{F}, (F_1 \vee F_2) \in \mathcal{F}, (F_1 \wedge F_2) \in \mathcal{F}$. The atomic formulas $\mathcal{AF}$ and these three logical operators are sufficient to write the usual comparisons and logical operators, which we introduce as notations: $\forall T_1, T_2 \in \mathcal{T}, \forall F_1, F_2 \in \mathcal{F}, (T > 0) \equiv \neg((-1 \cdot T) \geq 0), (T_1 \geq T_2) \equiv ((T_1 + (-1 \cdot T_2)) \geq 0), (T_1 > T_2) \equiv ((T_1 + (-1 \cdot T_2)) > 0), (F_1 \rightarrow F_2) \equiv (\neg F_1 \vee F_2)$ and $(T_1 = T_2) \equiv ((T_1 \geq T_2) \wedge (T_2 \geq T_1))$.

Example 2. $\mathrm{Exp}(X_1) \geq (3 \cdot \mathrm{Exp}(X_2))$ *is a formula.*

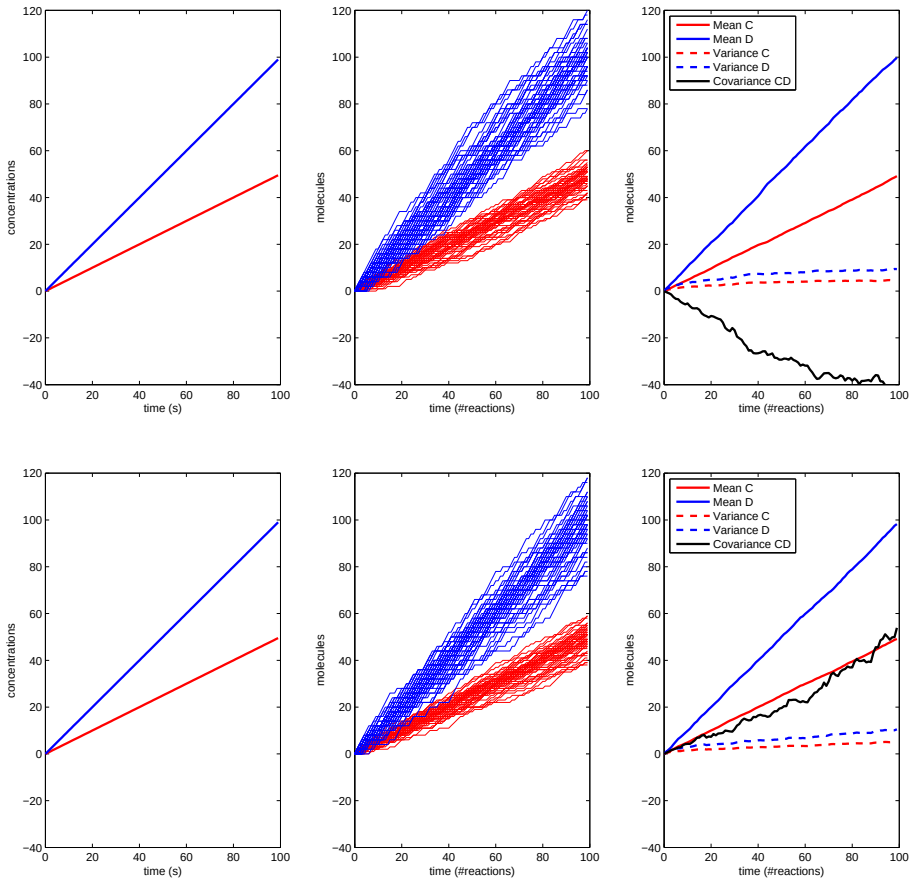

Figure 1: Differential and stochastic dynamics of model 1 (first row) and model 2 (second row). Red (resp. Blue) plots refer to quantities of C (resp. D). The first column depicts the solution to the differential equations derived from the law of mass actions. The second column depicts 50 runs of a stochastic simulation of the models (kinetic parameters equal to 1, $1000A$ and $1000B$ initially). The third column depicts the estimated mean, variance and covariance estimated from the simulations depicted in second column.

2.2 Semantics

Approximated moments in steady-state We want to define a relevant semantics of terms, that is a semantics corresponding to the moments (means, variances, covariances) of a stochastic process $(\vec{x}_k)_{k\in\mathbb{N}}$ that has a biologically correct distribution. We rely on a central limit theorem obtained in [Picard *et al.*, 2014] when considering *steady-state regime approximations*:

$$\frac{1}{\sqrt{k}}\left(\vec{x}_k - \left(\vec{x}_0 + kS^\top\vec{\mathfrak{p}}\right)\right) \xrightarrow[k\to+\infty]{\mathcal{D}} \mathcal{N}\left(\vec{0}, W(S,\vec{\mathfrak{p}})\right), \quad (4)$$

where $W(S,\vec{\mathfrak{p}}) = S^\top\left(\mathrm{diag}(\vec{\mathfrak{p}}) - \vec{\mathfrak{p}}\vec{\mathfrak{p}}^\top\right)S$ and k is the discrete-time variable. Here $\vec{\mathfrak{p}}$ is a m-dimensional probability vector named the *reaction probability vector* which represents the probabilities of triggering each reaction during the steady-state regime (that is when the distributions of reactants are stabilized). We denote by $\mathcal{P}_m$ the set of m dimensional probability vectors that is vectors $\vec{u}\in\mathbb{R}^m$ satisfying $\forall i\in\{1,\ldots,m\}, 0\le u_i\le 1$ and $\sum_{i=1}^m u_i = 1$. Therefore $\vec{\mathfrak{p}}\in\mathbb{P}_m$. Equation (4) provides us with asymptotic equivalents of the moments when $k\to\infty$:

$$\mathbb{E}x_k^a \sim_k x_0^a + k\sum_{j=1}^m s_{ja}\mathfrak{p}_j, \quad (5)$$

$$\mathbb{V}x_k^a \sim_k k\sum_{j=1}^m s_{ja}^2\mathfrak{p}_i - k\sum_{1\le j,l\le m} s_{ja}s_{la}\mathfrak{p}_j\mathfrak{p}_l, \quad (6)$$

$$\mathbb{C}ov(x_k^a, x_k^b) \sim_k k\sum_{j=1}^m s_{ja}s_{jb}\mathfrak{p}_j - k\sum_{1\le j,l\le m} s_{ja}s_{lb}\mathfrak{p}_j\mathfrak{p}_l, \quad (7)$$

where $u_k\sim_k v_k$ means the mathematical asymptotic equivalence of sequences, that is $u_k = v_k + o(v_k)$. Therefore the approximated first and second moments of $\vec{x}_k$ can be obtained when knowing the triplet $(S,\vec{x}_0,\vec{\mathfrak{p}})$. This motivates the following definition for the possible models of the formulas.

Definition 1. *A* context *is a pair* $C = (S,\vec{x}_0)$ *where* $\vec{x}_0\in\mathbb{Q}^n$

represents the initial quantities at the start of steady-sate regime and S is a $m \times n$ stoichiometry matrix. An interpretation *is a triplet $I = (S, \vec{x}_0, \vec{p})$, where $(S, \vec{x}_0)$ is a context, and $\vec{p} \in \mathcal{P}_m$ are reaction probabilities.*

Evaluation of terms When a context is given, the terms can be evaluated as multivariate polynomials with variables corresponding to the time k and reaction probabilities $\vec{p} = (p_i)_{0 \leq i \leq m}$. The evaluation of leaves when $\vec{p} := \vec{\mathfrak{p}}$ corresponds to the $\mathbb{R}[k]$ polynomial asymptotic expressions given in (5), (6) and (7).

Definition 2 (Evaluation of terms)**.** *The* evaluation $[T]_C$ *of a the term T in the context $C = (S, \vec{x}_0)$ is the polynomial $\mathbb{Q}[k, p_1, \ldots, p_m]$ defined by structural induction as*

$$[\mathrm{Exp}(X_a)]_C = x_0^a + k \sum_{j=1}^{m} s_{ja} p_j,$$

$$[\mathrm{Var}(X_a)]_C = k \left(\sum_{j=1}^{m} s_{ja}^2 p_i - \sum_{1 \leq j,l \leq m} s_{ja} s_{la} p_j p_l \right),$$

$$[\mathrm{Cov}(X_a, X_b)]_C = k \left(\sum_{j=1}^{m} s_{ja} s_{jb} p_j - \sum_{1 \leq j,l \leq m} s_{ja} s_{lb} p_j p_l \right),$$

$[c]_C = c$ *when c is a constant,* $[(\lambda \cdot T)]_C = \lambda \cdot [T]_C$, $[(T_1 + T_2)]_C = [T_1]_C + [T_2]_C$, $[(T_1 \times T_2)]_C = [T_1]_C [T_2]_C$.

The following proposition, stating that $[T]_C$ corresponds to the above asymptotic approximation of $\vec{x}_k$ when evaluated with $\vec{p} = \vec{\mathfrak{p}}$, justifies the definition of the semantics of terms.

Proposition 1. *Consider a reaction network with stoichiometry matrix S, initial state $\vec{x}_0$ and steady-state reaction probability vector $\vec{\mathfrak{p}}$ and a term T (i.e. a polynomial expression of the first and second moments). We denote by u_k the natural mathematical interpretation of T in terms of polynomial of expectancies, variances and covariances of $\vec{x}_k$, then $[T]_C(k, \vec{\mathfrak{p}}) \sim_k u_k$ when $k \to \infty$.*

Proof. (sketch) The proof is done by structural induction on T. □

Evaluation and models of formulas

Definition 3 (Evaluation of formulas)**.** *The* evaluation $[F]_C$ *of a the formula F in the context $C = (S, \vec{x}_0)$ is the subset of $\mathcal{P}_m$ defined by structural induction as*

$$[(T \geq 0)]_C = \{\vec{p} \in \mathcal{P}_m : \mathrm{dom}_k([T]_C) \geq 0\}, \qquad (8)$$

where $\mathrm{dom}_k(P) \in \mathbb{Q}[p_1, \ldots, p_m]$ is the dominant coefficient in k in the polynomial $P \in \mathbb{Q}[k, p_1, \ldots, p_m]$, $[\neg F]_C = \mathcal{P}_m \setminus [F]_C$, $[(F_1 \vee F_2)]_C = [F_1]_C \cup [F_2]_C$, $[(F_1 \wedge F_2)]_C = [F_1]_C \cap [F_2]_C$.

Therefore, the evaluation of an atomic formula $(T \geq 0)$ is the subset of probability vectors $\vec{p} \in \mathcal{P}_m$ such that $k \mapsto [T]_C(k, \vec{p}) \in \mathbb{Q}[k]$ is asymptotically non negative (since the asymptotic behavior of a polynomial is given by his monomial of highest degree).

Using this definition we are now able to define what are the models of a formula. An interpretation $I = (S, \vec{x}_0, \vec{p})$ is a *model* of a formula F, noted

$$I \models F, \quad \text{if } \vec{p} \in [F]_{(S, \vec{x}_0)}. \qquad (9)$$

A formula F is *valid*, noted $\models F$, if every interpretation is a model. A formula F is *valid in a context* $C = (S, \vec{x}_0)$, noted $C \models F$, if $\forall \vec{p} \in \mathcal{P}_m$, $(S, \vec{x}_0, \vec{p}) \models F$. A formula F is *satisfiable in a context* $C = (S, \vec{z}_0)$, if there exists $\vec{p} \in \mathcal{P}_m$ such that $(S, \vec{x}_0, \vec{p}) \models F$.
It follows that models of an atomic formula are triplets $(S, \vec{x}_0, \vec{p})$ such that the comparison is satisfied in the sense of the next proposition.

Proposition 2. *The interpretation $I = (S, \vec{z}_0, \vec{p})$ is a model of $F = (T \geq 0)$ if and only if*

$$\exists K \in \mathbb{N}, \quad \forall k \geq K, \quad [T]_{(S, \vec{z}_0)}(k, \vec{p}) \geq 0. \qquad (10)$$

Therefore, considering Proposition 1, an interpretation $I = (C, \vec{p})$ is a model of a comparison means that the comparison between the two polynomial expressions of moments is ultimately true in the framework of the steady-state approximation when $\vec{p} = \vec{\mathfrak{p}}$.

Proof. Denote $f(x) = [T]_C(x, \vec{p})$ for $x \in \mathbb{R}$. The function f is a polynomial in $\mathbb{Q}[x]$, so it has only a limited number of possible asymptotic behavior: either f is constant, or $\lim_{+\infty} f = +\infty$, or $\lim_{+\infty} f = -\infty$. If $I \models F$, then by definition $\mathrm{dom}_k([T]_C) \geq 0$ meaning that either f is a non negative constant or f is non-constant with positive dominant coefficient. In both cases (10) holds. Conversely, if (10) holds then either f is constant or $\lim_{+\infty} f = +\infty$, so $\mathrm{dom}_k([T]_C) \geq 0$, so $I \models F$. □

In our definitions of models, we pay attention to distinguish between valid formulas which are always true (for instance $(7 \geq 5)$ or $(\mathrm{Exp}(X_1) \geq 2\,\mathrm{Exp}(X_1))$) and formulas valid in a context, that is properties whose validity is a consequence of the topology and the stoichiometry of the considered reaction network. Valid properties in a context correspond to asymptotic properties that are true for all steady-state reaction probability vectors. Hence a reaction network can exhibit an asymptotic behavior F without having $C \models F$. However, if such an asymptotic property is observed in a presumed steady-state then the formula F must be satisfiable in the considered context. This last remark is very important because it allows one to *reject* a context $(S, \vec{x}_0)$, and especially to reject S, if a formula F coding for experimental observations of a presumed steady-state is not satisfiable in the considered context.

3 Deciding Satisfiability and Validity

We have defined the validity and satisfiability in a context of a formula in the previous section. The next step is to design algorithms for determining the validity or satisfiability of a given formula. The following lemma shows that both notions are in close relationship, so we can focus on the satisfiability problem.

Lemma 1. • *A formula F is valid in the context C if and only if $[F]_C = \mathcal{P}_m$.*

- *A formula F is satisfiable in the context C if and only if $[F]_C \neq \varnothing$.*
- *In the context C, a formula F is valid (resp. satisfiable) if and only if $\neg F$ is not satisfiable (resp. not valid).*

Theoretical Complexity We now demonstrate that the satisfiability problem is NP-hard by using a reduction from 3-SAT.

Proposition 3. *The following $\mathcal{F}$-SAT satisfiability problem is NP-hard.*
The $\mathcal{F}$-SAT problem
Instance: *n (number of chemical species), m (number of reactions), S ($n \times m$ stoichiometry matrix), $\vec{x}_0$ (initial quantities), F (a formula).*
Question: *Is there $\vec{p} \in \mathcal{P}_m$ such that $(S, \vec{x}_0, \vec{p}) \models F$?*

Therefore, there is no algorithm that can verify for an arbitrary reaction network the satisfiability (or the validity due to Lemma 1) of a formula in polynomial time unless P = NP.

Proof. The proof is obtained by polynomial time reduction from 3-SAT.
3-SAT decision problem [Garey and Johnson, 2002]
Instance: n (number of variables), a propositional formula in conjunctive normal form (CNF) $\varphi = \bigwedge_{i=1}^{r}(l_i^1 \vee l_i^2 \vee l_i^3)$, where $l_i^{1/2/3}$ are literals.
Question: Is there a valuation satisfying φ?
We provide a polynomial time reduction from 3-SAT. Consider a propositional formula in CNF $\varphi = \bigwedge_{i=1}^{r}(l_i^1 \vee l_i^2 \vee l_i^3)$ and denote by $\{x_1, \ldots, x_n\}$ the n variables of φ. From φ we build an associated formula of our logic $F = \bigwedge_{i=1}^{r}(G_i^1 \vee G_i^2 \vee G_i^3)$ where $G_i^q = (\exp(X_k) > 0)$ if $l_i^q = x_k$ and $G_i^q = \neg(\exp(X_k) > 0)$ if $l_i^q = \neg x_k$. Then we prove φ is satisfiable if and only if F is satisfiable.

- Let $v : \{x_1, \ldots, x_n\} \to \{\top, \bot\}$ be a valuation satisfying φ. We consider the following reaction network with n chemical species and $n+1$ reactions $\{R_0 : \varnothing \to \varnothing, R_i : \varnothing \to X_i (i = 1 \ldots n)\}$ with stoichiometry matrix S. Then we define the reaction probabilities as $p_k = 1/n$ if $v(x_k) = \top$, $p_k = 0$ if $v(x_k) = \bot$ and $p_0 = 1 - \sum_{k=1}^{n} p_k$. We also set initial conditions at zero $\vec{x}_0 = 0$. Let us consider the interpreation $I = (S, \vec{x}_0, \vec{p})$. Then $I \models (\exp(X_k) > 0) \Leftrightarrow p_k > 0 \Leftrightarrow v(x_k) = \top$ and $I \models \neg(\exp(X_k) > 0) \Leftrightarrow p_k \leq 0 \Leftrightarrow v(x_k) = \bot$. Consequently $I \models F$.
- Conversely, if $I = (S, \vec{x}_0, \vec{p}) \models F$ then we define a valuation as $v(x_k) = \top$ if $I \models (\mathrm{Exp}(X_k) > 0)$ and $v(x_k) = \bot$ if $I \models \neg(\mathrm{Exp}(X_k) > 0)$. By definition of $\models$ it follows that v satisfies φ.

□

An algorithm for deciding $\mathcal{F}$-SAT We have proven that deciding the satisfiability of a formula is NP-hard, nevertheless it is still interesting to design an algorithm for deciding this problem. Indeed, it is possible to find algorithms that are fast in practice but slow for a few specific reaction networks or formulas. We propose the following Algorithm 1 for determining $\exists ? \vec{p} \in \mathcal{P}_m,\ (S, \vec{x}_0, p) \models F$.

Algorithm 1: Deciding $\mathcal{F}$-SAT

Data: A context $C = (S, \vec{x}_0)$, a formula F
Result: $\vec{p}$ such that $(C, \vec{p}) \models F$ or UNSAT
Step 1: Convert F into disjunctive normal form (DNF);

$$F = \bigvee_{u=1}^{r} F_u = \bigvee_{u=1}^{r} (G_u^1 \wedge \cdots \wedge G_u^{m_u})$$

for $u = 1$ *to* r **do**
  Try Step 2: find $\vec{p} \in [F_u]_{(S,\vec{x}_0)}$;
  if $\vec{p}$ *is found* **then**
    return $\vec{p}$;
  end
end
return UNSAT;

In step 2, finding $\vec{p} \in [(T_u^q \geq 0)]_C$ (resp. $[\neg(T_u^q \geq 0)]_C$) corresponds to finding a solution $\vec{p}$ such that $(\mathrm{dom}_k[T]_{(S,\vec{x}_0)})(\vec{p}) \geq 0$ (resp. < 0), that is finding a solution to a polynomial inequality. Therefore, step 2 consists of finding a solution to a set of n_u polynomial constraints. Finding such a solution is decidable [Tarski, 1951] and can be performed by state-of-the-art model checking tools such as the SMT-solver `dReal` [Gao *et al.*, 2013].
The algorithm has two sources of complexity. First, converting F in DNF in step 1 can be computationally intensive since the size of the DNF can be exponential in the size of F and thus r may be large. This should not be a significant problem in usual cases since the formula F is not complex. Second, solving step 2, that is finding a solution to a system of polynomial constraints can be computationally intensive.

Terms without multiplications lead to quadratic constraints We have proposed a procedure for determining the satisfiability of a formula F based on solving sets of polynomial inequalities. Since general systems of polynomial inequalities can be difficult to solve we propose to consider a logical fragment of $\mathcal{F}$ by restricting terms to linear mathematical expressions of moments. Formally we define $\mathcal{T}_{\text{lin}} \subset \mathcal{T}$ in the same way as $\mathcal{T}$ but by removing the last induction rule : $\forall T_1, T_2 \in \mathcal{T},\ (T_1 \times T_2) \in \mathcal{T}$. We then define $\mathcal{F}_{\text{lin}} \subset \mathcal{F}$ with the same induction rules but using the terms in $\mathcal{T}_{\text{lin}}$.

Proposition 4. *Consider a context C and a term $T \in \mathcal{T}_{lin}$ then finding $\vec{p} \in \mathcal{P}_m,\ \vec{p} \in [(T \geq 0)]_C$ can be done by solving a numerical quadratic inequation in the variables (p_i).*

Proof. (sketch) The proof consists in demonstrating by structural induction on the terms that for all $T \in \mathcal{T}_{\text{lin}}$, the total degree for the variables (p_i) of $[T]_C$ is at most two. Indeed, the $\mathrm{Exp}(\cdot)$ leaves are polynomials of degree at most 1 for (p_i) , and the $\mathrm{Var}(\cdot)$ and $\mathrm{Cov}(\cdot, \cdot)$ leaves are polynomials of degree at most two. Then, summing terms and multiplication by a scalar do not increase the degree. □

As multiplication of terms is not used in the proof of Proposition 3, the satisfaction problem for $\mathcal{F}_{\text{lin}}$ is still NP-hard. However, using the algorithm described in the previous section may be simpler as the involved systems of constraints are quadratic. For instance, the constraints correspond to a second-order cone programming (SOCP) [Alizadeh and Goldfarb, 2003] problem which can be solved by interior point methods in tools such as `CPLEX` and `Gurobi`.

4 Example

Let us go back to the example of the introduction which was not possible to solve using classical fluxes based analysis. We consider that the biological experimental data are given by the *first row* of Figure 1. Also, as a consequence of the assumed steady-state, we consider that A and B are balanced, so their quantities do not change in average. Thus the data are encoded into the formula

$$F = (\text{Exp}(A) = 1000) \wedge (\text{Exp}(B) = 1000) \\ \wedge (\text{Exp}(D) \geq 2\,\text{Exp}(C)) \wedge (\text{Cov}(C, D) < 0). \quad (11)$$

Now we want to discriminate between the two reaction models, hence we introduce the two contexts $C_1 = (S_1, \vec{x}_0)$ and $C_2 = (S_2, \vec{x}_0)$ associated with each reaction network in order to check the satisfiability of F in both contexts. Here we know the initial conditions $\vec{x}_0 = (1000, 1000, 0, 0)$. F is already in DNF, so we directly derive the corresponding set of polynomial constraints using from semantics of the formulas.

atomic formulas	context C_1	context C_2
$(\text{Exp}(A) = 1000)$	$p_2 - p_1 = 0$	$p_2 - p_1 = 0$
$(\text{Exp}(B) = 1000)$	$p_1 - p_2 = 0$	$p_1 - p_2 = 0$
$(\text{Exp}(D) \geq 2\,\text{Exp}(C))$	$p_2 \geq p_1$	$0 \geq 0$
$(\text{Cov}(C, D) < 0)$	$-2p_1p_2 < 0$	$2p_1(1 - p_1) < 0$

As expected, the constraints are at most quadratic in $\vec{p}$ since there is no multiplication in the formula F. The first system of quadratic constraints admits the (unique) solution $\vec{p} = (1/2, 1/2)$ whereas the second system of constraints has no solution. Consequently, F is satisfiable in the context C_1 but not satisfiable in the context C_2

$$\exists \vec{p} \in \mathcal{P}_m, (C_1, \vec{p}) \models F \qquad \nexists \vec{p} \in \mathcal{P}_m, (C_2, \vec{p}) \models F.$$

So, the steady-state properties F cannot be obtained using the second reaction network $(model\ 2)$ which must me rejected.

5 Conclusion

We have introduced a logic whose syntax permits to express properties on the asymptotic first and second moments of the trajectories of a reaction network in steady-state. The semantics of formulas is obtained using a central limit theorem which provides us with analytical expressions of first and second moments. A model of a formula is a reaction network with initial quantities and a steady-state reaction probability vector such that the corresponding Gaussian asymptotic approximation satisfies the formula. When a formula encoding for experimental data is not satifiable in a given context, it means that the context and possibly the stoichiometry matrix is wrong. Thus, our logic provides a refutation of reaction networks based on the measurements of asymptotic first and second moments of the trajectories.

After introducing the logic, we have demonstrated that the $\mathcal{F}$-SAT problem is NP-hard. We provided an algorithm which relies on the DNF conversion and polynomial constraints solving. This open perspectives of improving the satisfiability test by using efficient constraints solving tools. Further work will focus on understanding which instances of $\mathcal{F}$-SAT can be solved in reasonable time. This includes a precise study of the practical complexity of the algorithm on various instances, that are pairs of biological models and datasets, with different sizes of reaction networks and different numbers and types of constraints.

Acknowledgments

This work was supported by the French National Research Agency via the investment expenditure program IDEALG (ANR-10-BTBR-02-11). The authors are grateful to Pr. Phillipe Codognet for interesting discussions about solving systems of polynomial constraints.

References

[Alizadeh and Goldfarb, 2003] Farid Alizadeh and Donald Goldfarb. Second-order cone programming. *Mathematical programming*, 95(1):3–51, 2003.

[Arkin *et al.*, 1998] Adam Arkin, John Ross, and Harley H McAdams. Stochastic kinetic analysis of developmental pathway bifurcation in phage λ-infected escherichia coli cells. *Genetics*, 149(4):1633–1648, 1998.

[Gao *et al.*, 2013] Sicun Gao, Soonho Kong, and Edmund M Clarke. dreal: An smt solver for nonlinear theories over the reals. In *Automated Deduction–CADE-24*, pages 208–214. Springer, 2013.

[Garey and Johnson, 2002] Michael R Garey and David S Johnson. *Computers and intractability*, volume 29. wh freeman, 2002.

[Gillespie, 1976] Daniel T Gillespie. A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. *Journal of computational physics*, 22(4):403–434, 1976.

[Gillespie, 2007] Daniel T Gillespie. Stochastic simulation of chemical kinetics. *Annu. Rev. Phys. Chem.*, 58:35–55, 2007.

[Helms, 2008] Volkhard Helms. *Principles of computational cell biology*. Wiley, 2008.

[McAdams and Arkin, 1997] HarleyH. McAdams and Adam Arkin. Stochastic mechanisms in gene expression. *Proceedings of the National Academy of Sciences*, 94(3):814–819, 1997.

[Orth *et al.*, 2010] Jeffrey D Orth, Ines Thiele, and Bernhard Ø Palsson. What is flux balance analysis? *Nature biotechnology*, 28(3):245–248, 2010.

[Picard *et al.*, 2014] Vincent Picard, Anne Siegel, and Jérémie Bourdon. Multivariate Normal Approximation for the Stochastic Simulation Algorithm: limit theorem and

applications. In *SASB - 5th International Workshop on Static Analysis and Systems Biology*, Munchen, Germany, 2014.

[Sandmann, 2008] Werner Sandmann. Discrete-time stochastic modeling and simulation of biochemical networks. *Computational biology and chemistry*, 32(4):292–297, 2008.

[Tarski, 1951] Alfred Tarski. A decision method for elementary algebra and geometry. *Rand report*, 1951.

[Wilkinson, 2009] Darren J Wilkinson. Stochastic modelling for quantitative description of heterogeneous biological systems. *Nature Reviews Genetics*, 10(2):122–133, 2009.